

Detection Theory on Random Graphs

Tom Mifflin
Advanced Mathematics Applications Division
Metron, Inc.
Reston, VA, U.S.A.
mifflin@metsci.com

Abstract - This paper presents some results in a theory of detection on random graphs. An Erdős-Rényi random graph serves as our noise model. Signals are represented by specific types of subgraphs embedded in the noise graph or other known structural characteristics. The paper begins with some known results about the expected number of subgraphs of a specific type in a random graph. This result is used to convince the reader of his likely poor intuition on which types of subgraphs will commonly appear in the noise graph. A detection problem called the prescribed subgraph problem is presented. A closed form calculation of the optimal detection statistic, i.e., the likelihood ratio, is the main result. Other results on detection theory for Erdős-Rényi random graphs are presented along with some results for other random graph models.

Keywords: Random graphs, Detection, Likelihood ratio.

1 Introduction

This paper was inspired by an attempt to model detections of surreptitious behavior from transactional data. In the link analysis community it is common to model transactions among individuals as links and nodes, respectively, in a graph. The usual situation for a link analysis problem is that most of the transactions are legitimate—very few transactions are threatening. Legitimate transactions constitute “transactional noise”. The goal is to detect those links and nodes that represent threatening activities or patterns in the midst of the noise transactions.

Discovering threat activity links and nodes is confounded by legitimate or noise transactions. This is because the threat activities are typically carried out in the midst of enormous amounts of legitimate transactions. Even when legitimate transactions are singleton events (i.e., not correlated with other transactions), in the aggregate they can result in structures of transactional data that are similar to the structure of the threat signatures, making it difficult to distinguish between the two.

Transactional “noise” is fundamentally different than classic models of noise in standard signal processing. Complex patterns, even those that resemble the threat patterns of interest can arise from simple noise alone. As the level of noise increases, the expected number of random occurrences of a particular pattern of transactions can grow exponentially, causing false alarm rates to become unmanageable. Thus, as in most detection problems, simply understanding the nature of the transactional signal is not sufficient. To successfully discriminate between signal and noise and to optimize the search for threat signatures, one must understand and model the transactional noise.

Even with a simple noise model, estimating the likelihood that a particular signature will arise from background noise can defy intuition. In Section 2, a standard random graph generation scheme is used to model transactional noise. A number of results from random graph theory are used to illustrate the counter-intuitive nature of threat detection in a noisy, transactional data environment and to contrast it with more classical detection problems.

In Section 3, we use the same random graph noise model to pose a simple detection problem. We present a closed-form expression for the likelihood ratio statistic in this context, which is a statistic that is an optimal detection statistic for a wide range of detection problems. The formula for computing the likelihood ratio statistic is then generalized to problems for which the transactions (links) have probabilities less than one of being observed in the data (i.e., partially-observed networks).

In Section 4, we move from detecting known patterns of transactions that correspond to particular subgraphs to cases in which there is no fixed known pattern of transactions. Instead, we consider a hypothesis test that evaluates whether the set of transactions constitute a single group that communicates at a uniform rate or at least two subpopulations that communicate at different rates. That is, the communicate rate within a given population is not the same as the communication rate across the populations. For this test, we present an explicit expression for the likelihood ratio statistic, and illustrate its calculation for various instantiations of the graph model.

2 Definition of the Erdős-Rényi Random Graph

This section presents the definition of the Erdős-Rényi random graph and a few well known results.

The purpose of this section is to highlight characteristics and phenomena of transactional noise using a simple random graph model. Our initial investigations will answer questions such as
1. What types of subgraphs rarely arise from background noise?

2. What types of subgraphs commonly arise from background noise?

3. How does the occurrence rate of a given subgraph vary with increasing noise?

4. Given a pattern graph, does it have an exploitable sub-pattern, that is, one that is unlikely to arise from noise? (When searching for instances of a pattern graph in a data graph, searching first for matches to the rarest sub-patterns will provide the greatest reduction in the search space and slow the combinatorial explosion of branch-based search algorithms.)

5. What are the key statistics to characterize a pattern graph or subgraph and to quantify the difficulty of a particular detection problem?

While there are several standard random graph models, this section restricts the discussion to a specific model. Let \( G(n,p) \) denote the model for generating a random graph among \( n \) entities in which each edge (of the \( n(n-1)/2 \) possible edges) is instantiated independently with probability \( p \). An instantiation of \( G(n,p) \) represents uncorrelated transactional noise.

In real applications, one might want to detect instances of a particular pattern of transactions that is believed to be indicative of threatening or illegal activity. In the context of this model, that goal translates into finding subgraphs of a particular type or pattern (to be mathematically rigorous, the goal is to find subgraphs of a particular isomorphism class). The pattern of transactions is analogous to a signature in a standard detection problem. A particular activity may have multiple signatures or indicators of its existence. In the context of this model, there could be several candidate subgraphs for which to search. The suitability of a candidate signature depends on the nature of the background noise transactions. For example, a subgraph that commonly arises from the background noise would be a poor choice for a subgraph signature.

The following theorem of Paul Erdös, one of the earliest results in random graph theory, provides a formula for computing a statistic that one might use to quantify the “rarity” of any subgraph of \( G(n,p) \).

**Theorem 1** (Erdös ([1], p. 218-9)). Let \( H \) be a subgraph with \( k \) vertices and \( e(H) \) edges. The expected number of subgraphs of type \( H \) in a random graph \( G = G(n,p) \) is

\[
E[X_H(G)] = \frac{n_k}{|\text{aut}(H)|} p^{e(H)},
\]

where \( n_k \) is the \( k^{th} \) falling factorial \( n(n-1) \ldots (n-k+1) \) and \( |\text{aut}(H)| \) is the size of the automorphism group of \( H \).

The automorphism group of \( H \), \( \text{aut}(H) \) is the set of permutations of vertices of \( H \) that preserve adjacency relations. The quotient \( n_k/|\text{aut}(H)| \) is the number of possible distinct subgraphs of type \( H \) among \( n \) vertices. Theorem 1 provides a closed form expression for the expected number of subgraphs of a given type in terms of the number of possible subgraphs, the number of subgraph edges and the edge probability.

Quantifying the “rarity” of a subgraph using the expected number from Theorem 1 leads to findings that are, perhaps, counter-intuitive. Figure 1 illustrates this with four simple subgraphs and the random graph \( G(n,p) \) with \( n=1000 \) and \( p=1/500 \).

**Figure 1.** Expected number of subgraphs for \( G(1000,0.002) \)

It may seem surprising that the expected number of “winged” 3-cycles is greater than the expected number of 3-cycles since each subgraph of the former type must contain the latter. The missing insight is that each triangle may be a subgraph of multiple winged triangles. For \( n = 1000 \) and \( p = 1/500 \), the expected number of wings on each triangle vertex is two. On average, we can expect eight winged triangles for each distinct triangle.

One might also guess (incorrectly) that the expected number of 60-cycles is less than the expected number of 3-cycles in an instantiation of \( G(1000,0.002) \) because of the 60-cycle’s greater number of edges and the improbability (just 1 in 500) of each edge. However, Erdős’ formula reveals that the expected number of 60-cycles is fifteen orders of magnitude greater than the number of 3-cycles. The insight that makes this result more intuitive is that the number of possible 60-cycles within a 1000-node graph, roughly \( 10^{177} \), is far greater than the number of possible 3-cycles, roughly \( 10^8 \).

It is worthwhile to briefly discuss the role of symmetry in the frequency or rarity of particular types of subgraphs. Recall that the automorphism group of a subgroup \( H \) is the set of vertex permutations that preserve adjacency relations in \( H \). For example, the automorphism group of a complete graph \( K_m \), \( \text{aut}(K_m) \), is the symmetric group \( S_m \) of size \( m! \). Likewise, \( \text{aut}(C_m) = D_m \), the dihedral group which has \( 2m \) elements.

The size of the automorphism group alone can greatly affect the frequency or rarity of a particular type of subgraph. In Figure 2, each subgraph has 75 edges and 75 vertices. Hence the differences in the expected values are due solely to the sizes of their automorphism groups.

For any finite group, \( G \), it is known that there exists a graph with \( G = \text{aut}(H_G) \). On the other hand, for large \( n \), most subgraphs of a graph of order \( n \) will have a trivial automorphism group. A good reference for graph automorphisms is [4].
Another measure of “rarity” is the probability that at least one subgraph of a given type arises from background noise. Bollobas proved that the existence or non-existence of a particular subgraph depends only upon its maximum density.

Definition 1: The maximum density of a graph $F$ is

$$m(F) = \max \left\{ \frac{e(H)}{v(H)} : H \subseteq F, v(H) > 0 \right\},$$

where $e(H)$ is the number of edges of $H$ and $v(H)$ is the number of vertices of $H$.

Theorem 2 ([2], p. 56): For an arbitrary graph $F$ with at least one edge,

$$\lim_{n \to \infty} P(G(n, p) \supseteq F) = \begin{cases} 0 & \text{if } p \ll \frac{1}{m(F)} \\ 1 & \text{if } p \gg \frac{1}{m(F)} \end{cases}$$

Theorem 2 says that there is a threshold probability for subgraphs. Figure 3 shows that this metric can also lead to counter-intuitive results. Given the large expected number of 60-cycles, it is not surprising that the probability of getting at least one 60-cycle from the background noise is nearly one. However, consider the subgraph consisting of a 60-cycle appended to a $K_4$. The expected number of occurrences is over a billion, yet the probability that at least one will occur is near zero.

$$H = \begin{array}{c}
\includegraphics[width=0.2\textwidth]{60-cycle}
\end{array} \quad E(H) = 10^{21}$$

$$H = \begin{array}{c}
\includegraphics[width=0.2\textwidth]{K4}
\end{array} \quad E(X_{H}) = 2 \times 10^{15}$$

$$H = \begin{array}{c}
\includegraphics[width=0.2\textwidth]{60-cycle+K4}
\end{array} \quad E(X_{H}) = 3 \times 10^{9}$$

$$P(X_{H} > 0) = \text{Near 1} \quad \text{Near 0} \quad \text{Near 0}$$

Figure 3. Expected number and probability of existence for three subgraphs of $G(1000, 0.002)$

The intuition behind this “paradox” is that a $K_4$ is unlikely to occur in any particular instantiation of $G(1000, 0.002)$, so a $K_4$ appended to a 60-cycle must also be unlikely. However, when a $K_4$ does occur, it is likely to share an edge with many 60-cycles, which occur in great abundance.
We seek a criterion for determining whether an observed graph \( J \) was generated according to \( GH(n,p) \) or \( G(n,p) \). That is, when does \( J \) constitute sufficient evidence to call a “detection” for the presence of a target? The optimal statistic on which to base this decision is the likelihood ratio

\[
\Lambda_H(J) = \frac{P(J \mid \text{target present})}{P(J \mid \text{no target})} = \frac{P(Z = J \mid Z \sim G_H(n,p))}{P(Y = J \mid Y \sim G(n,p))}.
\]

The following theorem is the first detection theory result in a structured, linked data environment.

**Theorem 3** (Mifflin, et al. [5]). Let \( G = G(n,p) \) and \( H \) denote the target graph. The likelihood ratio of an observed subgraph \( J \) is

\[
\Lambda_H(J) = \frac{X_H(J)}{E[X_H(G)]},
\]

where \( X_H(J) \) is the number subgraphs of type \( H \) in \( J \).

**Proof.** The denominator probability in equation (5) is

\[
P(J \mid \text{no target}) = p^{e(J)}(1 - p)^{\binom{n}{2} - e(J)}.
\]

The numerator probability of the likelihood ratio is

\[
P(J \mid \text{target present}) = \sum_{[\hat{H}] \subseteq H} P(J \mid \text{target is } \hat{H})P(\text{target is } \hat{H} \mid \text{target present})
\]

\[
= \sum_{[\hat{H}] \subseteq H} p^{e(\hat{H})}(1 - p)^{\binom{n}{2} - e(\hat{H})} \frac{1}{X_H(K_n)}
\]

The number of terms in the above sum is \( X_H(J) \), so

\[
P(J \mid \text{target present}) = p^{e(J)-e(\hat{H})}(1 - p)^{\binom{n}{2} - e(J)} \frac{X_H(J)}{X_H(K_n)}.
\]

Therefore,

\[
\Lambda_H(J) = \frac{X_H(J)}{p^{e(J)}X_H(K_n)} = \frac{X_H(J)}{E[X_H(G)]}.
\]

The numerator is the number of subgraphs of the evidence graph \( J \) that are isomorphic to \( H \), and the denominator is the expected number of subgraphs isomorphic to \( H \) that arise due to uncorrelated noise.

The denominator of equation (5) is the expected number of subgraphs of type \( H \) in an instance of the pure noise model \( G = G(n,p) \). This can be computed from the Erdös equation Theorem 1. The numerator will be, in general, much harder to calculate. Determining whether there are any subgraphs of type \( H \) contained in an observed graph \( J \) is an instance of the subgraph isomorphism problem, which is known to be NP-complete [3]. However, algorithms exist that can solve the vast majority of these problems extremely efficiently. Still, counting the number of distinct subgraphs of type \( H \) contained in \( J \) could prove to be challenging.

To complicate matters further, actual intelligence collection has some unreliability. For example, not every transaction will be observed in the data. This can be modeled by applying an “observability filter” to each instantiated random graph. Let \( G(n, p, q) \) and \( G_H(n, p, q) \) denote models for generating random graphs by first generating instances of \( G(n, p) \) or \( G_H(n, p) \), respectively, and then selecting each edge independently with probability \( q \). Theorem 4 generalizes Theorem 3 by deriving the likelihood ratio statistic \( \Lambda_q(J') \) for an observed graph \( J \) when all transactions are not observed:

\[
\Lambda_q(J') = \sum_{J' \subseteq J} w(J')\Lambda(J'),
\]

where the sum is taken over all supergraphs \( J' \) of \( J \) among the \( n \) vertices, \( \Lambda(J') \) is the likelihood ratio for discriminating between \( G_H(n, p, q) \) and \( G(n, p, q) \), and

\[
w(J') = \left( \frac{(1 - q)p}{1 - p} \right)^{e(J') - e(J)} \left( \frac{1 - pq}{1 - p} \right)^{\binom{n}{2} - e(J)}.
\]

Skokan provides an algorithm for enumerating the supergraphs \( J' \) of any subgraph \( J \) and proves that their weights \( w(J') \) sum to one [6].

Note that Theorem 4, while it does provide a likelihood ratio formula that is generalized to the context of partially observable transactional data, still assumes a simple model of uncorrelated background noise. The next section addresses this shortcoming.

4 Pattern-Free Detection

In the previous section, we looked at detecting known patterns of transactions that correspond to particular subgraphs. In this section we look at the case in which there is no fixed known pattern of transactions. Instead, we entertain the possibility that there exists at least two subpopulations that communicate at different rates. That is, the communication rate within a given population is not the same as the communication rate across the populations.
In the simplest case, we consider two groups of individuals. We can define another random graph process denoted by \( G(n_1, n_2, p_1, p_2, p_{12}) \) where

\[
\begin{align*}
    p_1 &= \text{link probability between group 1} \\
    p_2 &= \text{link probability between group 2} \\
    p_{12} &= \text{link probability between group 1 and group 2} \\
    n_1 &= \text{the number of individuals in group 1} \\
    n_2 &= \text{the number of individuals in group 2}
\end{align*}
\]

Given an observation graph \( J \), we would like to know if \( J \) was generated by \( G(n, p) \) or \( G(n_1, n_2, p_1, p_2, p_{12}) \). As before, the right statistic is the likelihood ratio,

\[
\Lambda(J) = \frac{P(J \mid G(n_1, n_2, p_1, p_2, p_{12}))}{P(J \mid G(n, p))}. \tag{8}
\]

It is useful to consider the two distinct sets of vertices of \( J \), \( V_1 \) and \( V_2 \), which correspond to the \( n_1 \) individuals in group 1 and \( n_2 \) individuals in group 2, respectively. Consider three subgraphs of \( J \), denoted by \( J_1, J_2, \) and \( J_{12} \). Sets \( J_1 \) and \( J_2 \) are the induced subgraphs on \( V_1 \) and \( V_2 \), respectively, and are illustrated in Figure 4 as blue and red nodes and edges, respectively. The bipartite graph \( J_{12} \) has vertex sets \( V_1 \) and \( V_2 \) in which the green edges are those in \( J \) going from \( V_1 \) to \( V_2 \).

![Figure 4. Decomposition of observed graph \( J \) into subgraphs \( J_1 \) (blue), \( J_2 \) (red) and \( J_{12} \) (green).](image)

The number of edges in each of these subgraphs is denoted by \( e(J_1), e(J_2), \) and \( e(J_{12}) \), respectively. We can expand the likelihood ratio statistic as

\[
\Lambda(J) = \frac{\binom{n_1}{2} p_1^{e(J_1)} p_{12}^{e(J_{12})} (1-p_1)^{\binom{n_1}{2}} (1-p_{12})^{e(J_{12})}}{\binom{n}{2} p_1^{e(J)} p_{12}^{e(J_{12})}}.
\]

Using the properties that \( e(J) = e(J_1) + e(J_2) + e(J_{12}) \), \( n = n_1 + n_2 \), and the identity

\[
\binom{n}{2} = \binom{n_1}{2} + \binom{n_2}{2} + n_1 n_2,
\]

we can expand the denominator of the likelihood ratio as

\[
\Lambda(J) = \frac{\binom{n_1}{2} p_1^{e(J_1)} p_{12}^{e(J_{12})} (1-p_1)^{\binom{n_1}{2}} (1-p_{12})^{e(J_{12})}}{\binom{n}{2} p_1^{e(J)} p_{12}^{e(J_{12})}}.
\]

Collecting common exponents, this expression simplifies to

\[
\Lambda(J) = \left( \frac{p_1}{\binom{n}{2}} \right)^{e(J_1)} \left( \frac{p_2}{\binom{n}{2}} \right)^{e(J_2)} \left( \frac{1-p_1}{\binom{n_1}{2}} \right)^{e(J_{12})} \left( \frac{1-p_2}{\binom{n_2}{2}} \right)^{e(J_{12})}.
\tag{9}
\]

The terms in equation (9) can collapse further depending upon the random graph model characteristics. For a common case such as \( p = p_1 \), the first and fourth terms in the product would be one, making the statistic independent of the value of \( e(J_1) \). Similarly, in the case that \( p_2 = p_{12} \), the statistic would not depend on the individual values of \( e(J_2) \) and \( e(J_{12}) \), but instead on just their sum \( e(J_2) + e(J_{12}) \).

### 4.1 Pattern-free detection examples

Figure 5 illustrates instantiations of \( G(n_1, n_2, p_1, p_2, p_{12}) \), all of which have \( n_1 = 21 \) and \( n_2 = 4 \). The upper left example in Figure 5 is an instance of \( G(21, 4, 0.5, 0.5, 0) \). The four vertices representing group 2 are easily visible near the top. The other three examples are instantiations of \( G(21, 4, 0.2, 0.19, 0.18) \), where group 2 is represented by the same four vertices. In order to calculate the likelihood ratio in equation (5), we assumed that \( p = p_1 \). That is, the null hypothesis assumes that everyone is in group 1.

![Figure 5. Multiple instantiations of \( G(n_1, n_2, p_1, p_2, p_{12}) \)](image)
In the upper left example in Figure 5, the likelihood ratio confirms that there are obviously distinct communities. This distinction is not as clear in the other examples. In the lower left example, the likelihood ratio favors the hypothesis that there is only one group present. In the lower right example, the ratio favors the hypothesis that there are two groups. This ambiguity is explained by connection probabilities, $p_1$, $p_2$ and $p_{12}$, being all nearly equal.

Fortunately, the situation is not hopeless. Rather than representing the total set of observations, what if each example represented the data collected over a single day. Collecting observation graphs over $N$ days, we can calculate the cumulative likelihood function,

$$\prod_{d=1}^{N} \Lambda(J_d),$$

where $J_d$ is the independent observation graph for day $d$.

Figure 6 shows the cumulative log-likelihood over a 100-day period for two cases: one in which there are two groups present, and one in which there is only one group present. Notice that over a small time interval, the results can be inconclusive, but over the entire 100-day timeline, the log-likelihood easily distinguishes the two cases.

### References


